



US Army Corps
of Engineers
Waterways Experiment
Station

DTIC
ELECTE
MAR 30 1993
S C D

Instruction Report W-93-1
February 1993

2

AD-A262 632



Water Operations Technical Support Program

Sampling Design Software User's Manual

by Robert F. Gaugush
Environmental Laboratory

WES

Approved For Public Release; Distribution Is Unlimited

Reproduced From
Best Available Copy

93-06527



93 3 30 063



Prepared for Headquarters, U.S. Army Corps of Engineers

20001016218

The contents of this report are not to be used for advertising, publication, or promotional purposes. Citation of trade names does not constitute an official endorsement or approval of the use of such commercial products.

Notice to Program Recipients

This program is furnished by the U.S. Government and is accepted and used by the recipient with the express understanding that the Government makes no warranty, expressed or implied, concerning the accuracy, completeness, reliability, usability, or suitability for any particular purpose of the information and data contained in this program or furnished in connection therewith, and the United States shall be under no liability whatsoever to any person by reason of any use made thereof. The program belongs to the Government. Therefore, the recipient further agrees not to assert any proprietary rights therein or to represent this program to anyone as other than a Government program.

All documents and reports conveying information obtained as a result of the use of the program by the recipient will acknowledge the U.S. Army Engineer Waterways Experiment Station, Corps of Engineers, Department of the Army, as the origin of the program. All such documentation will state the name and version of the program used by the recipient.



PRINTED ON RECYCLED PAPER

Water Operations Technical
Support Program

Instruction Report W-93-1
February 1993

Sampling Design Software User's Manual

by Robert F. Gaugush
Environmental Laboratory

U.S. Army Corps of Engineers
Waterways Experiment Station
3909 Halls Ferry Road
Vicksburg, MS 39180-6199

Final report

Approved for public release; distribution is unlimited

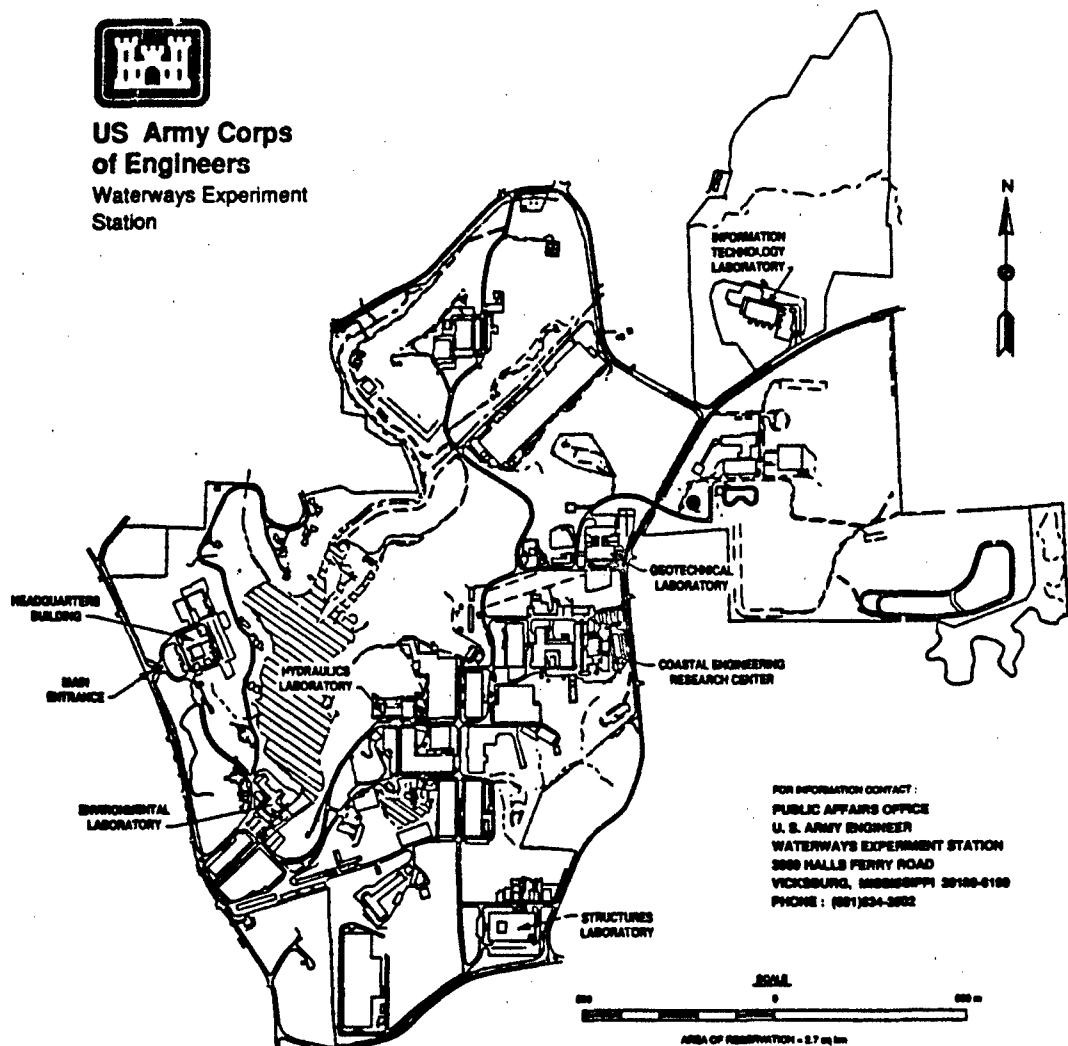
DTIC QUALITY INSPECTED 1

Prepared for U.S. Army Corps of Engineers
Washington, DC 20314-1000

Accession For	
NTIS CRA&I	<input checked="checked" type="checkbox"/>
DTIC TAB	<input type="checkbox"/>
Unannounced	<input type="checkbox"/>
Justification	
By	
Distribution /	
Availability Codes	
Dist	Avail and/or Special
A-1	



**US Army Corps
of Engineers**
Waterways Experiment
Station



FOR INFORMATION CONTACT:
PUBLIC AFFAIRS OFFICE
U. S. ARMY ENGINEER
WATERWAYS EXPERIMENT STATION
3600 HALLS FERRY ROAD
VICKSBURG, MISSISSIPPI 39180-6100
PHONE: (601) 334-3992

Waterways Experiment Station Cataloging-In-Publication Data

Gaugush, Robert F.

Sampling Design Software : user's manual / by Robert F. Gaugush ;
prepared for U.S. Army Corps of Engineers.

70 p. : ill. ; 28 cm. — (Instruction report ; W-93-1)

Includes bibliographical references.

1. Water quality — Measurement — Statistical methods — Computer
programs. 2. Sampling (Statistics) — Computer programs. 3. Statisti-
cal decision — Data processing. 4. Cluster analysis — Computer pro-
grams. I. United States. Army. Corps of Engineers. II. U.S. Army
Engineer Waterways Experiment Station. III. Title. IV. Title: Sampling
Design Software user's manual. V. Series: Instruction report (U.S. Army
Engineer Waterways Experiment Station) ; W-93-1

TA7 W34i no.W-93-1

Contents

Preface	v
1—Introduction	1
Background	1
Contents of the SDS Disk	1
Installation	3
Hardware Requirements	4
User Assistance	4
2—Decision Matrices	5
Program Execution	5
Data Entry	6
Error Messages	7
Documented Session	8
Example Output File	17
3—Variance Component Analysis	18
Data Set Preparation	18
Program Execution	22
Error Messages	23
Documented Session	23
Example Output File	29
4—Error Analysis	30
Data Set Preparation	30
Program Execution	31
Error Messages	32
Documented Session	32
Example Output File	43
5—Cluster Analysis	44
Data Set Preparation	45
Program Execution	46
Error Messages	46

Documented Session	47
Example Output File	59
References	61
Bibliography	62

Preface

This report was prepared by the Environmental Laboratory (EL) of the U.S. Army Engineer Waterways Experiment Station (WES), as part of the Water Quality Management for Reservoirs and Tailwaters Demonstration of the Water Operations Technical Support (WOTS) Program, sponsored by the U.S. Army Corps of Engineers (HQUSACE). Mr. Pete Juhle, HQUSACE, is Technical Monitor. The WOTS is managed under the Environmental Resources Research and Assistance Programs (ERRAP), Mr. J. Lewis Decell, WES, Manager. Dr. A. J. Anderson was Assistant Manager, ERRAP, for the WOTS program.

This report was prepared by Dr. Robert F. Gaugush of the Aquatic Processes and Effects Group (APEG), EL, under the direct supervision of Dr. Robert H. Kennedy, APEG, and under the general supervision of Mr. Donald L. Robey, Chief, Ecosystem Research and Simulation Division, EL, and Dr. John Harrison, Chief, EL.

At the time of publication of this report, Director of WES was Dr. Robert W. Whalin. Commander was COL Leonard G. Hassell, EN.

This report should be cited as follows:

Gaugush, Robert F. 1993. Sample Design Software User's Manual. Instruction Report W-93-1. Vicksburg, MS: U.S. Army Engineer Waterways Experiment Station.

1 Introduction

Background

The Sampling Design Software (SDS, Version 2.0) was developed as a companion to the Instruction Report "Sampling Design for Reservoir Water Quality Investigations" (Gaugush 1987). Four programs were developed to assist the user with problems with sampling design and its evaluation. The programs aid the decision-making process in sampling design through the use of decision matrices (the DECMATRX program). Sampling design evaluation is performed using variance component analysis (the VARCOM program), error analysis (the ERROR program), and cluster analysis (the CLUSTER program).

The purpose of this user's manual and the SDS disk provided with it is to assist the user in the implementation of these programs and is not intended to provide instruction on the assumptions and calculation methods of the statistical techniques used by these programs. The Bibliography presents a number of sources for basic statistics, sampling design, and more advanced statistical topics. The instruction report mentioned previously represents an introduction to the topic of sampling design. An introduction to statistics from a reservoir water quality perspective can be found in "Statistical Methods for Reservoir Water Quality Investigations" (Gaugush 1986).

Contents of the SDS Disk

A total of 39 files are provided on the SDS disk. The .EXE files are the compiled program files for DECMATRX, VARCOM, ERROR, and CLUSTER. These programs were developed and compiled using Turbo Pascal 5.5 (Borland International, Copyright 1984, 1989). The program files also have associated help files (files with an extension of .Hxx). Three example data sets are provided for the programs VARCOM, ERROR, and CLUSTER. These data sets are EG.VAR, EG.ERR, and EG.CLS, respectively.

Some files are required for all of the programs. The files with an extension of .BGI are graphics device drivers. Only one of these files will be used for any particular application, but all are provided for maximum compatibility with the numerous graphics cards to be found in personal computers (PC's). The files with an extension of .CHR are graphics character sets that are used in the introductory screens for each program. These files are supplied with the Turbo Pascal 5.5 compiler (Borland International, Copyright 1984, 1989).

The COLORS.DAT file is a short ASCII-format text file that is read by all of the programs to set the screen colors. If, after running the programs, you would like to change the screen colors, then simply edit this file. Notes on color selection are included in the file.

A complete listing of the files on the SDS disk is provided below:

Decision Matrices files:

DECMATRX.EXE - program file

DECMATRX.H01 - help files

DECMATRX.H02

DECMATRX.H03

DECMATRX.H04

DECMATRX.H05

Variance Component Analysis files:

VARCOM.EXE - program file

VARCOM.H01 - help files

VARCOM.H02

VARCOM.H03

EG.VAR - example data file

Error Analysis files:

ERROR.EXE - program file

ERROR.H01 - help files

ERROR.H02

ERROR.H03

ERROR.H04

ERROR.H05

EG.ERR - example data file

Cluster Analysis files:

CLUSTER.EXE - program file

CLUSTER.H00 - help files

CLUSTER.H01

CLUSTER.H02

CLUSTER.H03

CLUSTER.H04

CLUSTER.H05

CLUSTER.H06

CLUSTER.H07

CLUSTER.H08

CLUSTER.H09

EG.CLS - example data file

Files used for all programs:

ATT.BGI - graphics drivers

CGA.BGI

EGAVGA.BGI

HERC.BGI

IBM8514.BGI

PC3270.BGI

LITT.CHR - character sets

TRIP.CHR

COLORS.DAT - data file for setting screen colors

Installation

The SDS software will run from a single 360K 5.25-in. floppy disk (the software is supplied in this format), but performance will be improved considerably by installing the software on a hard disk drive.

To install the software on a hard disk:

- a. Create a subdirectory for the software

MD C:\SAMPLING

- b. Copy all files from the SDS disk to the new directory

CD \SAMPLING

COPY A: *.*

(The above examples assume that your C: drive is a hard disk and that the SDS disk is in drive A:)

Hardware Requirements

The SDS software has been tested on a number of different PC configurations. Testing has included 8088 (basic PC's), 80286 (AT types), and 80386 machines. Numeric co-processors are not required, but will be used if present. The CGA, EGA, VGA, and Hercules graphics drivers are supported.

User Assistance

Please contact:

Robert H. Kennedy, CEWES-ES-A
U.S. Army Engineer Waterways Experiment Station
3909 Halls Ferry Road
Vicksburg, MS 39180-6199

Telephone: (601) 634-3659

if you need assistance with the operation of the SDS software.

2 Decision Matrices

A decision matrix is an aid to the determination of sample size for multi-variable sampling programs and can be used for either simple random or stratified random sampling designs. The decision matrix is simply a tabular presentation that incorporates the factors necessary to determine sample size: (a) an estimate of the mean, (b) an estimate of the variability, (c) desired precision, (d) the acceptable probability of error, and (e) the costs associated with sampling. See Gaugush (1987) for a more complete discussion of determining sample size and the use of decision matrices.

Program Execution

To run the Decision Matrices program, simply type "decmatrx" at the DOS prompt. Be sure your default directory (i.e., the directory that you are in when you enter the above command) contains all of the files on the Sampling Design Software disk.

After the above command is entered, the program will prompt you for all of the necessary inputs. Program flow is as follows:

- a. Introductory screen.
- b. Prompt for output route - output may be routed to either the screen only or to a disk file as well as the screen (if disk file output is chosen, the program will prompt for a file name).
- c. Data entry.
- d. View output.
- e. Repeat analysis with new data.
- f. Exit program.

A documented session presented below provides a more complete view of the program flow.

Data Entry

DECMATRX is an interactive program and allows you to enter data during the execution of the program. Two data entry windows are used to (a) specify the parameters to be used by the program, and (b) enter estimates of the central tendency (i.e., the mean) and dispersion (i.e., the variance) of the variables to be sampled.

In the first data entry window, six fields are highlighted for input. (In the representations of the data entry windows shown below, highlighted fields are indicated by underlining the field.) In the first field enter the value (from 1 to 6) of the number of variables to be used in the decision matrix. The remaining fields are for the error probabilities and the levels of precision to be used in the analysis. Default values are provided for these fields, but they can be changed by entering the desired value in the respective field. Five possible values for the error probability are supported and are restricted to these values because of the method used to calculate the t statistic in the program. Values for precision can fall anywhere within the specified range of possible values. Generally, you will only need to specify the number of variables because the default values for error probability and precision provide a wide range of sample sizes.

```

      DECISION MATRIX
Number of variables (maximum of 6) : _
Error Probabilities : .05 .10 .20
      Default to   .05 .10 .20
      Possible values: .01 .05 .10 .20 .50
Levels of precision .10 .20
      Default to   .10 .20
      Range of possible values .01 TO .50
_____F1 - Help_____F2 - Continue_____

```

The arrow keys allow movement between the fields. The right and down arrows move the cursor to the next field while the left and up arrows move the cursor to the previous field. Typographical errors within a field can be corrected by using the backspace key to delete the error and then retyping the field. Errors can also be corrected after leaving the field that contains the error, but in this case the entire field must be retyped.

The second data entry window consists of four fields for each of the n variables specified in the first window. The example shown below assumes that the analysis is to be performed on three variables. As shown, a name, mean, coefficient of variation (C.V.), and cost must be specified for each variable. As before, the arrow keys allow for movement between the fields. Variable names can contain any characters (uppercase or lowercase, numbers may also be used), but blank spaces are not allowed in variable names. Decimal points are not required in the remaining fields but should be used for clarity. Values for the C.V.'s are expressed as a decimal fraction and not as a percentage. For example, the C.V. would be expressed as 0.50, not as 50.0 percent, for a variable with a mean of 50.0 and a standard deviation of 25.0.

DECISION MATRIX				
VARIABLE	NAME	MEAN	C.V.	UNIT COST
1	_____	_____	_____	_____
2	_____	_____	_____	_____
3	_____	_____	_____	_____

F1 - Help F2 - Continue

Error Messages

As the data are entered into the program, DECMATRIX checks for errors. The program checks the fields for number of variables, error probability, and precision for nonnumeric characters. If any are found, DECMATRIX will issue one of the following error messages:

INPUT ERROR: NUMBER OF VARIABLES INCORRECTLY ENTERED
 INPUT ERROR: ERROR PROBABILITY INCORRECTLY ENTERED
 INPUT ERROR: PRECISION INCORRECTLY ENTERED

The program also checks these same fields to determine if the values entered are within the range of values supported by the program. If any fall outside of the range of supported values, the program will issue one of the following messages:

INPUT ERROR: NUMBER OF VARIABLES IS OUT OF RANGE
 INPUT ERROR: ERROR PROBABILITY IS OUT OF RANGE
 INPUT ERROR: LEVEL OF PRECISION IS OUT OF RANGE

The second data entry window is also checked for errors. If a C.V. is less than or equal to zero, DECMATRX reports:

INPUT ERROR: C.V. <= 0

If a sampling cost is entered as a negative number, then the program issues the following error message:

INPUT ERROR: COST < 0

If any nonnumeric characters are entered for any of the means, C.V.'s, or costs, then one of the following messages will be displayed:

INPUT ERROR: MEAN INCORRECTLY ENTERED

INPUT ERROR: C.V. INCORRECTLY ENTERED

INPUT ERROR: COST INCORRECTLY ENTERED

Pressing any key after an error message has been reported will return the program to the data entry screen with the error. Correct the error and continue.

Documented Session

This example session with DECMATRX uses the following data:

<u>Variable</u>	<u>Mean</u>	<u>C.V.</u>	<u>Cost</u>
TP	95.	0.56	25.0
TN	1614.	0.28	25.0
CHLA	35.	0.52	25.0

The object of the analysis is to determine sample sizes and costs associated with sampling these three variables over an annual period. Sample sizes and costs for each variable are presented with respect to error probability and precision. The results of the analysis can be used to develop a sampling design within both statistical and financial constraints.

Entering the command "DECMATRIX" at the DOS prompt begins the program.

**Decision Matrices
Sampling Design Software – Version 2.0**

Developed by
Dr. Robert F. Gaugush
Environmental Laboratory
USAE Waterways Experiment Station

(Press any key to continue...)

Created using Turbo Pascal, Copyright Borland International 1984, 1989

After pressing any key, the program prompts for the output route.

Select output route

- 1) Screen only
- 2) Disk file

Enter value to continue...

F1 - Help

Press F1 for help.

Select output route

1) Screen only
2) Disk file

F1 - Help

Enter value to continue...

Help - Output routing

Output from the Decision Matrices program can be routed to a disk file as well as to the screen. If you select to output to a disk file, you will be prompted for a file name (paths can be included).

F2 - Continue

Press F2 to continue and clear the help window.

Select output route

1) Screen only
2) Disk file

F1 - Help

Enter value to continue...

Select 2 (disk file output). DECMATRX then prompts for the output file name. Use MATRIX.OUT for this session.

Input disk file name: matrix.out

DECMATRX then displays the first data entry window. (Underlined fields represent fields that will be highlighted on the PC screen).

```

      DECISION MATRIX
Number of variables (maximum of 6) : _
Error Probabilities : .05 .10 .20
      Default to   .05 .10 .20
      Possible values: .01 .05 .10 .20 .50

Levels of Precision  .10 .20
      Default to   .10 .20
      Range of possible values .01 TO .50

```

F1 - Help F2 - Continue

Press F1 for help.

DECISION MATRIX

Number of variables (maximum of 6) :

Error Probabilities : .05 .10 .20

Default to .05 .10 .20

Help - Data input

Enter data in each of the high-lighted fields. Default values exist for the error probabilities and the levels of precision. If these values are satisfactory then you only need to enter a value for the number of variables.

To move between fields: left or up arrow - previous field
 right or down arrow - next field

F1 - Help F2 - Continue

Press F2 to continue and clear the help window. Enter a "3" in the field for the number of variables.

Press F2 to continue and the program displays the second data entry window.

DECISION MATRIX

VARIABLE	NAME	MEAN	C.V.	UNIT COST
1	<u> </u>	<u> </u>	<u> </u>	<u> </u>
2	<u> </u>	<u> </u>	<u> </u>	<u> </u>
3	<u> </u>	<u> </u>	<u> </u>	<u> </u>

F1 - Help F2 - Continue

Press F1 for help.

DECISION MATRIX				
VARIABLE	NAME	MEAN	C.V.	UNIT COST
1	_____	_____	_____	_____
2	_____	_____	_____	_____

Help - Data input

Enter data in each of the high-lighted fields. Provide a name, mean, coefficient of variation, and sampling cost for each variable. The sampling costs are usually analytical costs per sample. If costs are not an issue, simply enter a 1 for the cost for each variable.

To move between fields: left or up arrow - previous field
 right or down arrow - next field

F1 - Help F2 - Continue F2 - Continue

Press F2 to continue and clear the help window. DECMATRX returns to the data entry window. Enter data to produce the screen shown below.

DECISION MATRIX				
VARIABLE	NAME	MEAN	C.V.	UNIT COST
1	TP	95.	0.56	25.
2	TN	1614.	0.28	25.
3	CHLA	35.	0.52	25.

F1 - Help F2 - Continue

When data entry is completed, press F2 to continue. The program displays sample sizes with respect to variable, error probability, and precision.

SAMPLE SIZE						
PRECISION:	0.10			0.20		
ERROR:	0.05	0.10	0.20	0.05	0.10	0.20
VARIABLE						
TP	123	87	53	33	23	14
TN	33	23	14	10	7	4
CHLA	106	75	46	28	20	12

F1 - Help F2 - Exit F3 - Costs

Press F1 for help.

SAMPLE SIZE						
PRECISION:	0.10			0.20		
ERROR:	0.05	0.10	0.20	0.05	0.10	0.20
VARIABLE						
TP	123	87	53	33	23	14
TN	33	23	14	10	7	4
CHLA	106	75	46	28	20	12

Help - Sample sizes

Sample sizes are provided for each combination of variable, error probability, and precision.

F2 - Continue

F1 - Help F2 - Exit F3 - Costs

Press F2 to continue and clear the help window. Press F3 to see the costs window.

COST						
PRECISION:	0.10			0.20		
ERROR:	0.05	0.10	0.20	0.05	0.10	0.20
VARIABLE						
TP	3075	2175	1325	825	575	350
TN	825	575	350	250	175	100
CHLA	2650	1875	1150	700	500	300

F1 - Help F2 - Exit F3 - Sample size

Press F1 for help.

COST						
PRECISION:	0.10			0.20		
ERROR:	0.05	0.10	0.20	0.05	0.10	0.20
VARIABLE						
TP	3075	2175	1325	825	575	350
TN	825	575	350	250	175	100
CHLA	2650	1875	1150	700	500	300

Help - Sampling costs

Sampling costs are provided for each combination of variable, error probability, and precision.

F2 - Continue

F1 - Help F2 - Exit F3 - Sample size



Example Output File

DECISION MATRIX

INPUT DATA

ERROR PROBABILITIES : 0.10 0.20
 LEVELS OF PRECISION : 0.05 0.10 0.20

VARIABLE	MEAN	C.V.	UNIT COST
TP	9.500E+01	5.600E-01	2.500E+01
TN	1.614E+03	2.800E-01	2.500E+01
CHLA	3.500E+01	5.200E-01	2.500E+01

SAMPLE SIZE

PRECISION:		0.10		0.20			
ERROR:		0.05	0.10	0.20	0.05	0.10	0.20
VARIABLE							
TP	123	87	53	33	23	14	
TN	33	23	14	10	7	4	
CHLA	106	75	46	28	20	12	

COST

PRECISION:		0.10		0.20			
ERROR:		0.05	0.10	0.20	0.05	0.10	0.20
VARIABLE							
TP	3075	2175	1325	825	575	350	
TN	825	575	350	250	175	100	
CHLA	2650	1875	1150	700	500	300	

3 Variance Component Analysis

Variance component analysis is a technique for quantifying the sources of variability in the data resulting from a given sampling design. The analysis results in the determination of each design component's contribution to the overall variance. Based on these results, sampling effort allocated to a given component of the design could be reduced or eliminated. See Winer (1971) for a comprehensive treatment of variance component analysis.

Data Set Preparation

The VARCOM program requires that input data sets be prepared prior to its use (i.e., data input during the program is not available). Data sets can be prepared with most text editors and word processing software. The data sets may contain only ASCII characters and none of the special characters used by most word processors for formatting. If you use a word processor to generate your data sets, be sure to save the files in DOS or ASCII format.

Data in VARCOM input files are organized into four groups:

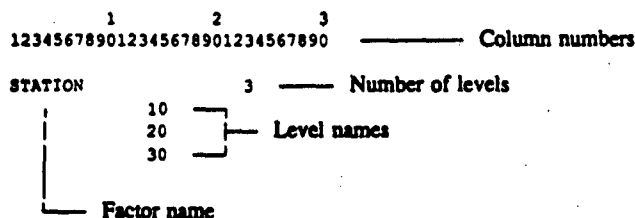
- Group 1 - title
- Group 2 - problem size identifiers
- Group 3 - factor and level information
- Group 4 - data records

An example data set, EG.VAR, is provided on the SDS distribution diskette and is shown below:

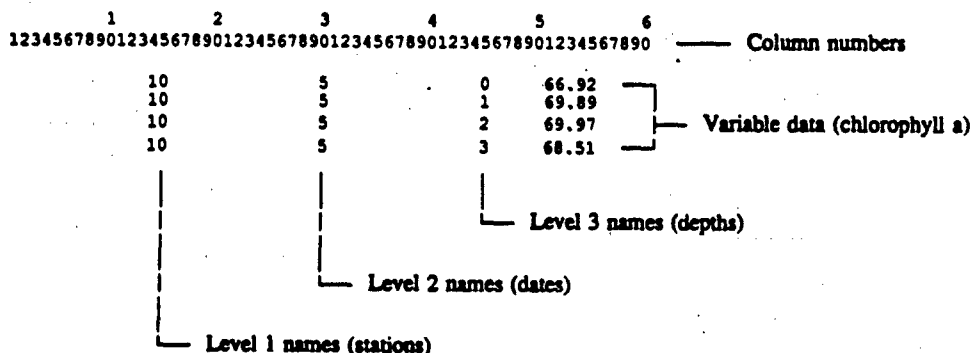
EAU GALLE - CHLOROPHYLL - MAY 1981				Data Group 1	
3 24				Data Group 2	
STATION	3				
10					
20					
30					
DAY	2				
5					
19				Data Group 3	
DEPTH	4				
0					
1					
2					
3					
10	5	0	66.92		
10	5	1	69.89		
10	5	2	69.97		
10	5	3	68.51		
10	19	0	4.77		
10	19	1	6.23		
10	19	2	4.38		
10	19	3	3.88		
20	5	0	46.13		
20	5	1	39.85		
20	5	2	44.17		
20	5	3	46.45	Data Group 4	
20	19	0	3.37		
20	19	1	3.38		
20	19	2	6.11		
20	19	3	4.71		
30	5	0	57.28		
30	5	1	48.00		
30	5	2	59.71		
30	5	3	58.39		
30	19	0	3.50		
30	19	1	3.70		
30	19	2	8.64		
30	19	3	6.47		

Data Group 1 consists of a single line specifying a title for the data set (maximum of 60 characters). Data Group 2 is a single line with two items. The first is the number of factors in the data set (VARCOM allows a maximum of three factors), and the second indicates the number of observations in Data Group 4. Data Group 3 names the factors, specifies the number of levels for each factor, and provides the name for each of the levels. A maximum of 100 levels is supported by VARCOM. In the example data set, three factors are specified in Data Group 2. The three factors used in the example data set are STATION, DAY, and DEPTH. STATION has three levels (10, 20, and 30) which means that three stations were sampled. DAY has two levels (samples were taken on the 5th and the 19th of May). Depth has four levels (samples were taken at 1-m intervals from the surface to 3 m). Data Group 4 lists the value of the variable to be analyzed (chlorophyll *a* in the example data set) for each combination of the factors. For example, at station 10 on the 5th of May at a depth of 1 m, the chlorophyll *a* concentration was 69.89 $\mu\text{g/l}$ (second line of Data Group 4).

VARCOM requires that the data in Data Groups 3 and 4 be placed in specific columns. A portion of Data Group 3 with column identifiers is shown below.



A factor name can have a maximum of 20 characters and must begin in column 1 (i.e., factor names must be left-justified). Separate the factor name and the number of its levels by one blank space. Therefore, the value for the number of levels should begin in column 22 or greater. A level name (in the following row) can have a maximum of 15 characters and must end in column 15 (i.e., all level names must be right-justified). A portion of Data Group 4 with column identifiers is shown below.



Level 1 names must end in column 15, level 2 names end in column 30, and level 3 names end in column 45. At least one blank column must separate the last level name from the variable data.

A data set for a two factor variance component analysis would appear as follows:

EAU GALLE - CHLOROPHYLL - MAY 1981

2 24

STATION		3	
	10		
	20		
	30		
DAY		2	
	5		
	19		
	10	5	66.92
	10	5	69.89
	10	5	69.97
	10	5	68.51
	10	19	4.77
	10	19	6.23
	10	19	4.38
	10	19	3.88
	20	5	46.13
	20	5	39.85
	20	5	44.17
	20	5	46.45
	20	19	3.37
	20	19	3.38
	20	19	6.11
	20	19	4.71
	30	5	57.28
	30	5	48.00
	30	5	59.71
	30	5	58.39
	30	19	3.50
	30	19	3.70
	30	19	8.64
	30	19	6.47

Note that multiple observations for combinations of levels are allowed. In the above data set, there are four observations for each combination of station and day. It also important to note that the order of lines in Data Group 4 is not important. The above data set could be just as correctly specified as:

EAU GALLE - CHLOFOPHYLL - MAY 1981

2 24

STATION		3	
	10		
	20		
	30		
DAY		2	
	5		
	19		
	10	5	66.92
	10	5	69.89
	10	5	69.97
	10	5	68.51
	20	5	46.13
	20	5	39.85
	20	5	44.17
	20	5	46.45
	30	5	57.28
	30	5	48.00
	30	5	59.71
	30	5	58.39
	10	19	4.77
	10	19	6.23
	10	19	4.38
	10	19	3.88
	20	19	3.37
	20	19	3.38
	20	19	6.11
	20	19	4.71
	30	19	3.50
	30	19	3.70
	30	19	8.64
	30	19	6.47

As long as the level names and the variable data on each line are placed in the proper position, then the lines of Data Group 4 can be arranged in any convenient order. A one factor data set would appear as follows:

EAU GALLE - CHLOROPHYLL - MAY 1981

1 24	
DAY	2
5	
19	
5	66.92
5	69.89
5	69.97
5	68.51
19	4.77
19	6.23
19	4.36
19	3.88
5	46.13
5	39.85
5	44.17
5	46.45
19	3.37
19	3.38
19	6.11
19	4.71
5	57.28
5	48.00
5	59.71
5	58.39
19	3.50
19	3.70
19	8.64
19	6.47

Suggestion: use an extension of .VAR for VARCOM data files. This will distinguish them from other data files.

Program Execution

To run the Variance Component Analysis program, simply type "var-com" at the DOS prompt. Be sure your default directory (i.e., the directory that you are in when you enter the above command) contains all of the files on the Sampling Design Software disk.

After the above command is entered, the program will prompt you for all of the necessary inputs. Program flow is as follows:

- a. Introductory screen.
- b. Prompt for output route - output may be routed to either the screen only or to a disk file as well as the screen (if disk file output is chosen, the program will prompt for a file name).
- c. Prompt for input file name.
- d. View output.
- e. Repeat analysis with new data.

f. Exit program.

A documented session presented below provides a more complete view of the program flow.

Error Messages

After prompting for the input and output file names, VARCOM performs an error check on the input data set. If the data set specifies more than three factors for the analysis, the program reports:

ERROR: NUMBER OF FACTORS EXCEEDS MAX. FACTORS

If the number of levels for any of the factors exceeds 100, the following error message is reported:

ERROR: NUMBER OF LEVELS FOR FACTOR 1
EXCEEDS THE MAX. NUMBER OF LEVELS

If the number of observations is greater than 3,500, VARCOM reports:

ERROR: NUMBER OF OBSERVATIONS EXCEEDS MAXIMUM

If, for any factor, the number of level names does not agree with the names listed, the program provides the following error message:

ERROR: LEVEL ID NOT FOUND

VARCOM terminates after reporting any of the above error messages. Edit the input data file and run the program again.

Documented Session

This example session with VARCOM uses the EG.VAR data set provided on the SDS distribution diskette. These data were derived from studies conducted on Eau Galle Reservoir in west-central Wisconsin. The data set has three factors: STATION, DAY, and DEPTH. STATION has three levels (stations 10, 20, and 30), DAY has two levels (the 5th and 19th of May), and DEPTH has four levels (depths of 0, 1, 2, and 3 m).

The object of the analysis is to determine the distribution of the variance in chlorophyll *a* among the three factors. If all of the factors account for a significant fraction of the variance in chlorophyll *a*, then the sampling design is efficient. If, on the other hand, one or two of the factors account for most of the variance, then the sampling effort could be

reduced. The sampling design could be modified to include only those factors that explain the majority of the variance.

Entering the command "VARCOM" at the DOS prompt begins the program.

**Variance Component Analysis
Sampling Design Software – Version 2.0**

Developed by
Dr. Robert F. Gaugush
Environmental Laboratory
USAE Waterways Experiment Station

(Press any key to continue...)

Created using Turbo Pascal, Copyright Borland International 1984, 1989

After pressing any key, the program prompts for the output route.

Select output route

- 1) Screen only
- 2) Disk file

F1 - Help

Enter value to continue...

Press F1 for help.

Select output route

- 1) Screen only
- 2) Disk file

F1 - Help Enter value to continue...

Help - Output routing

Output from the Variance Component Analysis program can be routed to a disk file as well as to the screen. If you select to output to a disk file, you will be prompted for a file name (paths can be included).

F2 - Continue

Press F2 to continue and clear the help window.

Select output route

- 1) Screen only
- 2) Disk file

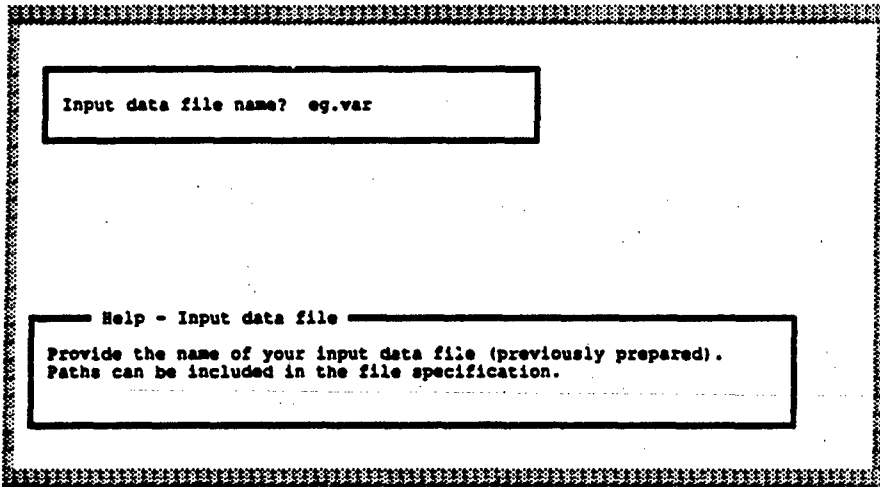
F1 - Help Enter value to continue...

Select 2 (disk file output). VARCOM then prompts for the output file name. Use EG.OUT for this session.



Disk file name for output: eg.out

The program then prompts for the input file name. Use EG.VAR for this session.



Input data file name? eg.var

Help - Input data file

Provide the name of your input data file (previously prepared).
Paths can be included in the file specification.

VARCOM then displays the results of the variance component analysis.

VARIANCE COMPONENT ANALYSIS			
EAU GALLE - CHLOROPHYLL - MAY 1981			
SOURCE	DF	SS	MS
STATION	2	6.30E+02	3.15E+02
DAY	1	1.58E+04	1.58E+04
DEPTH	3	4.52E+01	1.51E+01
ERROR	17	6.91E+02	4.07E+01
CORRECTED TOTAL	23	1.72E+04	
VARIANCE COMPONENT		ESTIMATE	PERCENT TOTAL
VAR(STATION)	3.43E+01	2.47
VAR(DAY)	1.31E+03	94.90
VAR(DEPTH)	-4.27E+00	< .01
VAR(ERROR)		4.07E+01	2.94
F1 - Help F2 - Exit			

Press F1 for help.

VARIANCE COMPONENT ANALYSIS			
EAU GALLE - CHLOROPHYLL - MAY 1981			
SOURCE	DF	SS	MS
STATION	2	6.30E+02	3.15E+02
Help - Variance component analysis			
<p>Output is divided into two sections. The upper section of the window provides the output of an n-way analysis of variance. The "Source" column lists the sources of variability within the data set. The "DF" column provides the degrees of freedom for each of the sources. The sum of squares and the mean square error are given in the "SS" and "MS" columns, respectively. The lower section of the output lists the variance component estimates and the relative contribution of each source to the overall variance.</p>			
F2 - Continue			
F1 - Help F2 - Exit			

Press F2 to continue and clear the help window.

VARIANCE COMPONENT ANALYSIS			
EAU GALLE - CHLOROPHYLL - MAY 1981			
SOURCE	DF	SS	MS
STATION	2	6.30E+02	3.15E+02
DAY	1	1.58E+04	1.58E+04
DEPTH	3	4.52E+01	1.51E+01
ERROR	17	6.91E+02	4.07E+01
CORRECTED TOTAL	23	1.72E+04	
VARIANCE COMPONENT		ESTIMATE	PERCENT TOTAL
VAR(STATION)	3.43E+01	2.47
VAR(DAY)	1.31E+03	94.90
VAR(DEPTH)	-4.27E+00	< .01
VAR(ERROR)		4.07E+01	2.94
F1 - Help F2 - Exit			

The variance component analysis indicates that most of the variance (almost 95 percent) is explained by sampling date (the DAY factor). For this data set, sampling stations and dates account for less than 3 percent of the total variance. Press F2 to exit.

Repeat program with new data? (Y or N) N

After finishing the analysis, you can repeat the program with a new data set or exit the program.

Example Output File

VARIANCE COMPONENT ANALYSIS

EAU GALLE - CHLOROPHYLL - MAY 1981

Title

SOURCE	DF	SS	MS	
STATION	2	6.30E+02	3.15E+02	N-way analysis of variance
DAY	1	1.58E+04	1.58E+04	
DEPTH	3	4.52E+01	1.51E+01	
ERROR	17	6.91E+02	4.07E+01	
CORRECTED TOTAL	23	1.72E+04		

VARIANCE COMPONENT		ESTIMATE	PERCENT TOTAL	
VAR(STATION))	3.43E+01	2.47	Variance component estimates
VAR(DAY))	1.31E+03	94.90	
VAR(DEPTH))	-4.27E+00	< .01	
VAR(ERROR)		4.07E+01	2.94	

4 Error Analysis

Error analysis is a statistical technique that can be used to improve an existing sampling design that uses the observed distribution of variance to redefine the sampling design. The results of the error analysis are used to redistribute samples to the existing strata to produce the minimum variance about the mean. The technique can be applied to the data of a stratified sampling design or to the data from a simple random or a systematic sample that has been subjected to poststratification (i.e., defining strata a posteriori). See Gaugush (1987) for a more detailed description of stratified sampling and the use of error analysis.

Data Set Preparation

The ERROR program requires that input data sets be prepared prior to its use (i.e., data input during the program is not available). Data sets can be prepared with most text editors and word processing software. The data sets may contain only ASCII characters and none of the special characters used by most word processors for formatting. If you use a word processor to generate your data sets, be sure to save the files in DOS or ASCII format.

Data in ERROR input files are organized into four groups:

- Group 1 - title
- Group 2 - problem size identifier
- Group 3 - strata weights
- Group 4 - data records

An example data set, EG.ERR, is provided on the SDS distribution diskette and is shown below:

```

EAU GALLE - 1981 - STATION 20 _____ Data Group 1
4 _____ Data Group 2
1 .167 _____ Data Group 3
2 .334 _____
3 .167 _____
4 .332 _____
1 7.8748E+01 _____
1 1.6735E+02 _____
1 4.2722E+01 _____
1 4.3925E+00 _____
2 5.8933E+00 _____
2 2.7610E+01 _____
2 2.9570E+01 _____
2 5.7273E+01 _____
2 4.2378E+01 _____
2 4.0602E+01 _____
2 5.5306E+01 _____
2 6.5534E+01 _____ Data Group 4
2 5.2158E+01 _____
3 3.1465E+01 _____
3 2.4320E+01 _____
3 4.0684E+01 _____
3 3.1248E+01 _____
4 1.3363E+01 _____
4 1.8966E+01 _____
4 1.0322E+01 _____
4 2.8420E+00 _____
4 3.5075E+00 _____
4 8.2300E+00 _____
4 2.8575E+01 _____
4 2.5618E+01 _____

```

Data Group 1 consists of a single line for the title of the data set (maximum of 60 characters). Data Group 2 also is a single line that specifies the number of strata in the data set. The ERKOR program supports a maximum of 25 strata. Data Group 3 specifies the strata numbers and weights. The strata numbers must be in numerical order and start with 1. The strata weights must sum to 1.00. At least one blank space must separate the stratum number and stratum weight in Data Group 3. Data Group 4 lists the observations of the sample data set consisting of the stratum number and the value of the variable (separated by at least one blank space). (Note: Although the example data set uses the computer representation of scientific notation (i.e., 2.5618E+01 is the computer form of 2.5618×10^1) for the data values, this is not required. These numbers could have been entered in a more typical decimal notation.)

Suggestion: use an extension of .ERR for ERROR data files. This will distinguish them from other data files.

Program Execution

To run the Error Analysis program, simply type "error" at the DOS prompt. Be sure your default directory (i.e., the directory that you are in when you enter the above command) contains all of the files provided on the Sampling Design Software disk.

After the above command is entered, the program will prompt you for all of the necessary inputs. Program flow is as follows:

- a. Introductory screen.
- b. Prompt for output route - output may be routed to either the screen only or to a disk file as well as the screen (if disk file output is selected, the program will prompt for a disk file name).
- c. Prompt for input file name.
- d. View output.
- e. Repeat analysis with new data.
- f. Exit program.

A documented session presented below provides a more complete view of program flow.

Error Messages

After prompting for the input and output file names, ERROR performs an error check on the input data set. If the data set specifies more than 25 strata for the analysis, the program reports:

ERROR : NUMBER OF STRATA EXCEEDS MAXIMUM

If the strata weights do not sum to 1.00, the following error message is reported:

ERROR : WEIGHTS DO NOT SUM TO 1.00

ERROR reports the following message if any of the strata have less than three observations:

ERROR : LESS THAN 3 SAMPLES IN STRATUM 1

Documented Session

This example execution of ERROR uses the EG.ERR data set provided on the SDS distribution diskette. These data were derived from studies conducted on Eau Galle Reservoir in west-central Wisconsin. Composite epilimnetic samples for chlorophyll *a* were taken at approximately 2-week intervals at Station 20 (a station located at the deepest part of the lake).

The data were stratified a posteriori into four strata: spring, summer, fall, and winter. The strata were defined as follows: "1" for spring - April and May (61 days), "2" for summer - June, July, August, and September (122 days), "3" for fall - October and November (61 days), and "4" for winter - December, January, and February (121 days). Strata weights were calculated by dividing the number of days in the stratum by 365.

The object of the analysis is to determine if the sampling design can be improved through the use of a stratified design using an optimal allocation of samples to the strata. Error analysis calculates the error variance associated with existing distribution of samples and determines an optimal distribution based on the observed variance among strata. If the existing and the optimal distribution of samples are considerably different, the sampling design can be improved by adopting the optimal distribution.

Entering the command "ERROR" at the DOS prompt begins the program.

Error Analysis

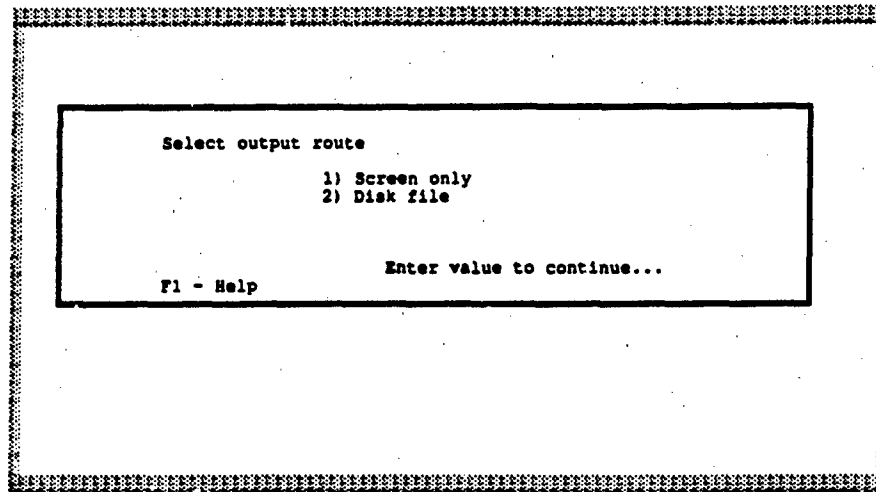
Sampling Design Software – Version 2.0

**Developed by
Dr. Robert F. Gaugush
Environmental Laboratory
USAE Waterways Experiment Station**

(Press any key to continue...)

Created using Turbo Pascal, Copyright Borland International 1984, 1989

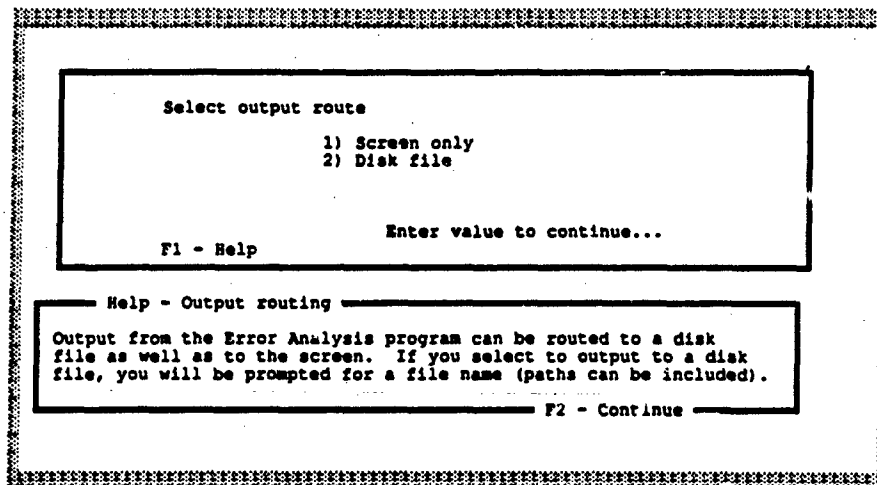
After pressing any key, the program prompts for the output route.



A screenshot of a terminal window showing a menu titled "Select output route". The menu lists two options: "1) Screen only" and "2) Disk file". Below the options, it says "Enter value to continue...". In the bottom left corner, it says "F1 - Help". The entire menu is enclosed in a rectangular box with a dotted border.

```
Select output route  
1) Screen only  
2) Disk file  
  
Enter value to continue...  
F1 - Help
```

Press F1 for help.

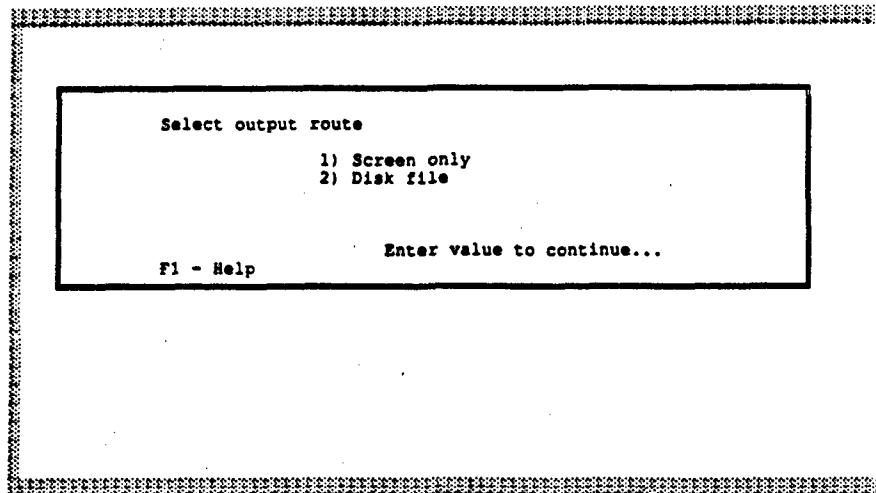


A screenshot of a terminal window showing the same "Select output route" menu as above. Below the menu, there is a help box titled "Help - Output routing" which contains text explaining that output can be routed to a disk file or the screen, and that selecting a disk file will prompt for a file name. The help box is also enclosed in a rectangular box with a dotted border. In the bottom right corner of the help box, it says "F2 - Continue".

```
Select output route  
1) Screen only  
2) Disk file  
  
Enter value to continue...  
F1 - Help
```

```
Help - Output routing  
Output from the Error Analysis program can be routed to a disk  
file as well as to the screen. If you select to output to a disk  
file, you will be prompted for a file name (paths can be included).  
F2 - Continue
```

Press F2 to continue and clear the help window.

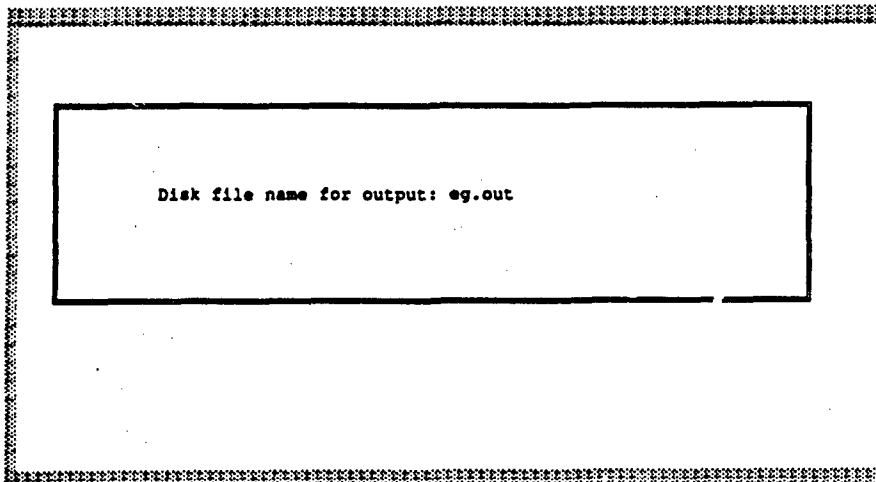


Select output route

- 1) Screen only
- 2) Disk file

F1 - Help Enter value to continue...

Select 2 (disk file output). ERROR then prompts for the output file name.
Use EG.OUT for this session.



Disk file name for output: eg.out

The program then prompts for the input file name. Use EG.ERR for this session.

Input data file name? eg.err

Help - Input data file

Provide the name of your input data file (previously prepared).
Paths can be included in the file specification.

ERROR then displays the statistics for the stratified sample.

EAU GALLE - 1981 - STATION 20

STRATIFIED SAMPLE STATISTICS

MEAN	3.62E+01
VARIANCE	1.85E+02
ERROR VARIANCE	3.97E+01

F1-Help F2-Sample Stat F3-Strata Stat F4-Analysis F5-Exit

Press F1 for help.

```
EAU GALLE - 1981 - STATION 20

STRATIFIED SAMPLE STATISTICS

      MEAN      3.62E+01
    VARIANCE    1.85E+02
  ERROR VARIANCE 3.97E+01

Help - Stratified sample statistics
Statistics (mean, variance, and error variance) for the stratified
sample.

F2 - Continue

F1-Help  F2-Sample Stat  F3-Strata Stat  F4-Analysis  F5-Exit
```

Press F2 to continue and clear the help window.

```
EAU GALLE - 1981 - STATION 20

STRATIFIED SAMPLE STATISTICS

      MEAN      3.62E+01
    VARIANCE    1.85E+02
  ERROR VARIANCE 3.97E+01

F1-Help  F2-Sample Stat  F3-Strata Stat  F4-Analysis  F5-Exit
```

Press F3 to see the strata statistics.

EAU GALLE - 1981 - STATION 20					
STRATA STATISTICS					
STRATUM	N	MEAN	VARIANCE	ERROR VARIANCE	
1	4	7.33E+01	4.85E+03	1.21E+03	
2	9	4.18E+01	3.42E+02	3.80E+01	
3	4	3.19E+01	4.51E+01	1.13E+01	
4	8	1.39E+01	9.34E+01	1.17E+01	

F1-Help F2-Sample Stat F3-Strata Stat F4-Analysis F5-Exit

Press F1 for help.

EAU GALLE - 1981 - STATION 20					
STRATA STATISTICS					
STRATUM	N	MEAN	VARIANCE	ERROR VARIANCE	
1	4	7.33E+01	4.85E+03	1.21E+03	
2	9	4.18E+01	3.42E+02	3.80E+01	
3	4	3.19E+01	4.51E+01	1.13E+01	
4	8	1.39E+01	9.34E+01	1.17E+01	

Help - Strata statistics

Statistics (number of samples, mean, variance, and error variance) for each of the sampled strata.

F2 - Continue

F1-Help F2-Sample Stat F3-Strata Stat F4-Analysis F5-Exit

Press F2 to continue and clear the help window. The screen returns to the strata statistics. Press F4 to see the results of the error analysis.

EAU GALLE - 1981 - STATION 20			
ERROR ANALYSIS			
STRATUM	% VARIANCE	% N	% OPTIMUM
1	85.3	16.0	52.5
2	10.7	36.0	27.9
3	0.8	16.0	5.1
4	3.2	32.0	14.5
VARIANCE WITH EXISTING DESIGN			3.97E+01
VARIANCE WITH OPTIMAL DESIGN			1.96E+01
F1-Help F2-Sample Stat F3-Strata Stat F4-Analysis F5-Exit			

Press F1 for help.

EAU GALLE - 1981 - STATION 20			
ERROR ANALYSIS			
STRATUM	% VARIANCE	% N	% OPTIMUM
1	85.3	16.0	52.5
Help - Error analysis			
<p>The %Variance column gives the relative contribution of each stratum to the overall stratified sample variance. The %N column shows how the samples were distributed among the strata. Using the observed distribution of variance among strata (the %Variance column), error analysis suggests an optimal distribution of samples among the strata (the %Optimum column). The reported "Variance with optimal design" is the error variance that would result if the optimal design was adopted for future sampling (if conditions do not dramatically change over time).</p>			
F2 - Continue			
F1-Help F2-Sample Stat F3-Strata Stat F4-Analysis F5-Exit			

Press F2 to continue and clear the help window.

EAU GALLE - 1981 - STATION 20			
ERROR ANALYSIS			
STRATUM	% VARIANCE	% N	% OPTIMUM
1	85.3	16.0	52.5
2	10.7	36.0	27.9
3	0.8	16.0	5.1
4	3.2	32.0	14.5
VARIANCE WITH EXISTING DESIGN			3.75E+01
VARIANCE WITH OPTIMAL DESIGN			1.93E+01
F1-Help F2-Sample Stat F3-Strata Stat F4-Analysis F5-Exit			

At any time during the program you can switch between the output windows. Press F2 to return to the sample statistics screen.

EAU GALLE - 1981 - STATION 20	
STRATIFIED SAMPLE STATISTICS	
MEAN	3.62E+01
VARIANCE	1.85E+02
ERROR VARIANCE	3.97E+01
F1-Help F2-Sample Stat F3-Strata Stat F4-Analysis F5-Exit	

Press F4 to return to the error analysis screen.

EAU GALLE - 1981 - STATION 20			
ERROR ANALYSIS			
STRATUM	% VARIANCE	% N	% OPTIMUM
1	85.3	16.0	52.5
2	10.7	36.0	27.9
3	0.8	16.0	5.1
4	3.2	32.0	14.5
VARIANCE WITH EXISTING DESIGN			3.97E+01
VARIANCE WITH OPTIMAL DESIGN			1.96E+01
F1-Help F2-Sample Stat F3-Strata Stat F4-Analysis F5-Exit			

The results of the error analysis indicate that the error variance could be reduced to less than 50 percent ($19.6/39.7 = 0.494$) of its observed value by using the optimal design. The optimal design consists of a redistribution of samples to place more samples in highly variable strata and less samples in strata with less variability. The spring stratum (stratum 1) accounts for over 85 percent of the observed variance (% Variance column), but only 16 percent (% N column) of the samples were allocated to this stratum. The optimal design would allocate just over 52 percent (% Optimum column) of the samples to this stratum. The winter stratum (stratum 4) accounts for only 3 percent of the observed variance, but 32 percent of the sampling effort was allocated to this stratum. The optimal design suggests that only about 15 percent of the samples should be dedicated to this stratum.

Press F5 to exit.

Repeat program with new data? (Y or N)

At this point you may choose to either run ERROR on another data set or exit from the program.

Example Output File

ERROR ANALYSIS

EAU GALLE - 1981 - STATION 20 _____ Title

STRATIFIED SAMPLE STATISTICS

MEAN	3.62E+01
VARIANCE	1.85E+02
ERROR VARIANCE	3.97E+01

Statistics for the entire stratified sample

STRATA STATISTICS

STRATUM	N	MEAN	VARIANCE	ERROR VARIANCE
1	4	7.33E+01	4.85E+03	1.21E+03
2	9	4.18E+01	3.42E+02	3.80E+01
3	4	3.19E+01	4.51E+01	1.13E+01
4	8	1.39E+01	9.34E+01	1.17E+01

Statistics for each
of the strata

ERROR ANALYSIS

STRATUM	% VARIANCE	% N	% OPTIMUM
1	85.3	16.0	52.5
2	10.7	36.0	27.8
3	0.8	16.0	5.1
4	3.2	32.0	14.5

Error analysis

VARIANCE WITH EXISTING DESIGN	3.97E+01
VARIANCE WITH OPTIMAL DESIGN	1.96E+01

5 Cluster Analysis

Cluster analysis is a multivariate classification technique that may be used to group or identify similar objects or entities. In a data analysis situation (rather than a sampling design evaluation), cluster analysis may be used to group a set of reservoirs according to their trophic state or by the composition of their phytoplankton. The use of cluster analysis in a typical data analysis mode can be found in Gaugush (1986). For the purposes of sampling design evaluation, cluster analysis can be used to identify and possibly reduce redundancies in the sampling design. The use of cluster analysis for this type of application is described more completely in Gaugush (1987).

In the evaluation of a sampling design, cluster analysis can be used to examine the quality of the information being provided by elements of the sampling design. In cluster analysis these elements are referred to as "entities" and may be sampling stations, dates, and/or the strata used in a stratified sampling design. The analysis begins with each entity in its own cluster and proceeds to join similar clusters until all of the entities are in a single cluster. The object, when used to evaluate a sampling design, is to determine if all of the elements of the design are providing independent information. For example, assume that data have been collected for twelve stations in a reservoir and a cluster analysis of the data indicates that the data fall into four clusters each represented by three stations. This implies that some of the stations are redundant (they are supplying essentially the same information). If the sampling program were to be continued (as in a monitoring program), the results of the cluster analysis could be used to reduce sampling effort. Sampling only 1 of the 3 stations from each cluster would result in the use of 4 stations rather than 12.

The CLUSTER program can be used to identify redundancies in sampling programs and suggest ways in which to reduce sampling effort in future studies. CLUSTER uses one of three clustering methods (average linkage, centroid, or Ward's method) to cluster the data; outputs a tabular "history" of the clustering; and produces a dendrogram of the clustering.

Data Set Preparation

The Cluster Analysis program requires that input data sets be prepared prior to its use (i.e., data input during the program is not available). Data sets can be prepared with most text editors and word processing software. The data sets may contain only ASCII characters and none of the special characters used by most word processors for formatting. If you use a word processor to generate your data sets, be sure to save the files in DOS or ASCII format.

Data in CLUSTER input files are organized into four groups:

- Group 1 - title
- Group 2 - problem size identifiers
- Group 3 - entity names
- Group 4 - data records

An example data set, EG.CLS, is provided on the SDS distribution diskette and is shown below:

EAU GALLE	_____	Data Group 1
5 3	_____	Data Group 2
STA10	} _____	Data Group 3
STA20		
STA30		
STA50		
STA60		
.069 1.507 44.129	} _____	Data Group 4
.078 1.503 43.144		
.068 1.473 41.155		
.068 1.427 33.800		
.070 1.487 46.068		

CLUSTER does not require strict positioning of data in specific columns, but it does have two simple requirements: (a) each line must start in column 1, and (b) multiple items on a single line must be separated by one blank space. Data Group 1 consists of a single line specifying a title for the data set (maximum of 60 characters). Data Group 2 is a single line with two items. The first is the number of entities in the data set, and the second indicates the number of variables to be used. The CLUSTER program can handle a maximum of 50 entities with a maximum of 10 variables. Data Group 3 provides the names of the entities (one line for each of the entities specified in Data Group 2). Each name can have a maximum of 20 characters. In the example data set, the entities are water quality sampling stations in Eau Galle Reservoir. Data Group 4 lists the data for the variables (one line for each entity and in the same order) to be used in the cluster analysis. In the example data set, these variables are total phosphorus, total nitrogen, and chlorophyll *a* concentrations (from left to right).

Suggestion: use an extension of .CLS for CLUSTER data files. This will distinguish them from other data files.

Program Execution

To run the Cluster Analysis program, simply type "cluster" at the DOS prompt. Be sure your default directory (i.e., the directory that you are in when you enter the above command) contains all of the files provided on the Sampling Design Software disk.

After the above command is entered, the program will prompt you for all of the necessary inputs. Program flow is as follows:

- a. Introductory screen.
- b. Prompt for input file name.
- c. Prompt for output file name.
- d. Prompt for clustering method.
- e. View output.
- f. Exit program.

A documented session presented below provides a more complete view of program flow.

Error Messages

After prompting for the input and output file names, CLUSTER performs an error check on the input data set. If the data set specifies either more than 50 entities or more than 10 variables in Data Group 2, CLUSTER outputs the following:

```
ERROR IN INPUT FILE
```

```
EITHER NUMBER OF ENTITIES  50 OR  
NUMBER OF VARIABLES  10
```

```
EDIT INPUT FILE AND BEGIN AGAIN
```

After displaying the error message the program terminates.

CLUSTER also performs an error check on Data Group 4. If the standard deviation of any of the variables is zero, CLUSTER outputs the following:

ERROR IN DATA

STANDARD DEVIATION FOR VARIABLE j IS ZERO

THIS MEANS THAT VARIABLE j IS THE SAME FOR
ALL ENTITIES AND WILL SERVE NO PURPOSE IN THE
CLUSTER ANALYSIS - DELETE THE VARIABLE FROM THE
INPUT FILE AND BEGIN AGAIN

As the error message states, a variable without variance (standard deviation equal to zero) does not add information to the cluster analysis. After displaying the error message, the program terminates.

Documented Session

This example execution of CLUSTER uses the EG.CLS data set provided on the SDS distribution diskette. These data were derived from studies conducted on Eau Galle Reservoir in west-central Wisconsin. The entities are five water quality stations within the reservoir. Stations 10 and 50 (STA10 and STA50) are littoral stations located in two different coves. Station 40 (STA40) is an inlet station. Station 30 (STA30) is located over the old river channel, and Station 20 (STA20) is located over the deepest portion of the pool. These stations were routinely sampled, and the data in Group 4 of EG.CLS are station means for total phosphorus, total nitrogen, and chlorophyll *a* in the epilimnion (0 - 3 m) for one growing season (April - September).

The object of the analysis is to determine if any of the stations are redundant. If two or more stations are supplying the same information, the possibility exists for reducing the number of stations. Reducing the number of stations brings about the obvious reduction in costs without reducing the information derived from the sampling program.

Entering the command "CLUSTER" at the DOS prompt begins the program.

Cluster Analysis Sampling Design Software – Version 2.0

Developed by
Dr. Robert F. Gaugush
Environmental Laboratory
USAE Waterways Experiment Station

<Press any key to continue...>

Created using Turbo Pascal, Copyright Borland International 1984, 1989

After pressing any key, the program prompts for the input file name.
For this session enter EG.CLS.

Input data file name? eg.cls

Provide the file name of your data file. Paths are accepted.

CLUSTER then prompts for the output file name. Use **EG.OUT** for this session.

```
Output data file name? eg.out

Provide a file name of your output data file. Paths are accepted.
```

At this point **CLUSTER** prompts for the method to be used in the cluster analysis. Help windows are available by pressing **F1**, **F2**, **F3**, or **F4**.

```
CLUSTERING METHOD:  AVERAGE LINKAGE (A)
                   CENTROID (C)
                   WARDS (W)

Enter choice of method...

F1 - General help
Specific help: F2 - Avg linkage F3 - Centroid F4 - Wards
```


Press F1 and the following is displayed.

```
CLUSTERING METHOD:  AVERAGE LINKAGE (A)
                   CENTROID (C)
                   WARDS (W)

Help - Clustering methods

Three methods (average linkage, centroid, and Wards) are available
to use to cluster the data. Select a method by entering the letter
associated with the desired method.

F2 - Continue
```

Press F2 to continue, and the help window is removed.

```
CLUSTERING METHOD:  AVERAGE LINKAGE (A)
                   CENTROID (C)
                   WARDS (W)

Enter choice of method...

F1 - General help
Specific help: F2 - Avg linkage F3 - Centroid F4 - Wards
```

Press F2 for the Average Linkage help window.

CLUSTERING METHOD: AVERAGE LINKAGE (A)
CENTROID (C)
WARDS (W)

Help - Clustering method: Average linkage

This clustering method has been found (along with Wards method) to be one of the more robust approaches to clustering data. In average linkage the distance between two clusters is the average distance between pairs of observations, one in each cluster. This method tends to produce clusters with small variance and is somewhat biased toward producing clusters with the same variance.

F2 - Continue

Pressing F2 (continue) again would remove the help window and restore the method selection screen. For the sake of brevity, assume F2 was pressed followed by F3 for the Centroid help window.

CLUSTERING METHOD: AVERAGE LINKAGE (A)
CENTROID (C)
WARDS (W)

Help - Clustering method: Centroid

This clustering method uses the distance between the centroids or means of the clusters. This method is more robust to the presence of outliers in the data than either the average linkage or Wards methods. In other respects, the centroid method may not perform as well as the other two methods.

F2 - Continue

Again assume F2 was pressed to return to the method selection screen and then F4 was selected to bring up the help window on Wards method.

CLUSTERING METHOD: AVERAGE LINKAGE (A)
CENTROID (C)
WARDS (W)

Help - Clustering method: Wards

This clustering method, although robust, tends to join clusters with a small number of observations and is biased toward producing clusters with generally the same number of observations. This method is also sensitive to the presence of outliers in the data.

F2 - Continue

Press F2 to return to the method selection screen.

CLUSTERING METHOD: AVERAGE LINKAGE (A)
CENTROID (C)
WARDS (W)

Enter choice of method...

F1 - General help

Specific help: F2 - Avg linkage F3 - Centroid F4 - Wards

Press A to select the average linkage method.

```
File: eg.out
```

```
Cluster Analysis
```



```
EAU GALLE
```


ID NUMBER	ENTITY
1	STA10
2	STA20
3	STA30
4	STA50
5	STA60


```
Average Linkage Method used for clustering
```

```
F1-Help F2-Exit Movement Keys: Home, End, PgUp, PgDn, Up and Down Arrows
```

At this point the cluster analysis is complete and you can view your output file (in this case EG.OUT as indicated in the first line). The cursor movement keys (Home, End, Page Up, Page Down, up arrow, and down arrow as indicated on the last line) allow you to browse through the output file. Press Page Down.

File: eg.out			
Stage	Clusters Joined		Distance
1	1	5	6.080E-01
2	1	3	1.219E+00
3	1	2	3.191E+00
4	1	4	6.000E+00

The distances are segmented into the following classes for the Linear dendrogram

CLASS	LOWER BOUND	UPPER BOUND
1	6.080E-01	8.237E-01
2	8.237E-01	1.039E+00
3	1.039E+00	1.255E+00
4	1.255E+00	1.471E+00
5	1.471E+00	1.686E+00

F1-Help F2-Exit Movement Keys: Home, End, PgUp, PgDn, Up and Down Arrows

The next 20 lines of the output file are displayed. Press Page Down again.

File: eg.out		
6	1.686E+00	1.902E+00
7	1.902E+00	2.118E+00
8	2.118E+00	2.333E+00
9	2.333E+00	2.549E+00
10	2.549E+00	2.765E+00
11	2.765E+00	2.981E+00
12	2.981E+00	3.196E+00
13	3.196E+00	3.412E+00
14	3.412E+00	3.628E+00
15	3.628E+00	3.843E+00
16	3.843E+00	4.059E+00
17	4.059E+00	4.275E+00
18	4.275E+00	4.490E+00
19	4.490E+00	4.706E+00
20	4.706E+00	4.922E+00
21	4.922E+00	5.137E+00
22	5.137E+00	5.353E+00
23	5.353E+00	5.569E+00
24	5.569E+00	5.784E+00
25	5.784E+00	6.000E+00

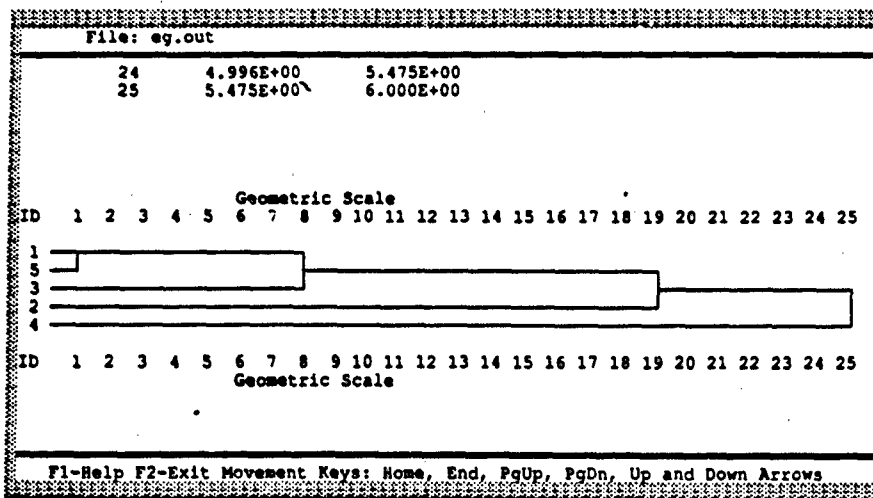
F1-Help F2-Exit Movement Keys: Home, End, PgUp, PgDn, Up and Down Arrows

Again the display moves 20 lines down. Press Home.

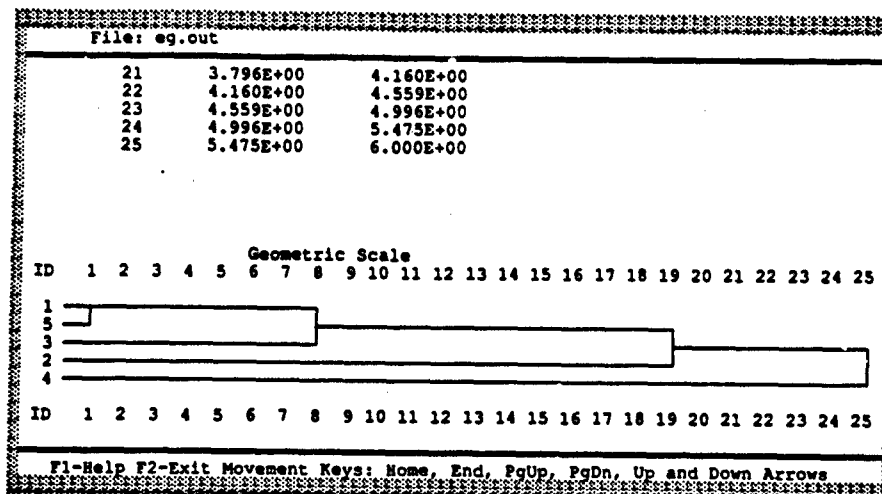
File: eg.out		
Cluster Analysis		
EAU GALLE		
ID	NUMBER	ENTITY
1		STA10
2		STA20
3		STA30
4		STA50
5		STA60
Average Linkage Method used for clustering		

F1-Help F2-Exit Movement Keys: Home, End, PgUp, PgDn, Up and Down Arrows

The display returns to the top of the output file. Press End.



The display moves to the bottom of the output file. Press the up arrow three times.



The display moves up three lines. The other movement keys operate in a similar manner. Press F1 for help.

```

File: eg.out
21 3.796E+00 4.160E+00
22 4.160E+00 4.559E+00
23 4.559E+00 4.396E+00
24 4.996E+00 5.475E+00
25 5.475E+00 6.000E+00
Help - Cluster analysis output

Short descriptions of various portions of the output are available.

F3 - Entity and ID numbers
F4 - Stages of clustering
F5 - Distances
F6 - Dendrogram
F7 - Linear vs. geometric scales for the dendrogram

F2 - Continue
F1-Help F2-Exit Movement Keys: Home, End, PgUp, PgDn, Up and Down Arrows
  
```

A help menu window is displayed over the output file. Press F3.

```

File: eg.out
21 3.796E+00 4.160E+00
22 4.160E+00 4.559E+00
Help - Entity listing

This section lists the ID numbers that have been assigned to the
entities in the data set. Entities can be stations, dates, depths,
reservoirs, etc. This listing will be necessary to interpret the
dendrogram.

F2 - Continue

F4 - Stages of clustering
F5 - Distances
F6 - Dendrogram
F7 - Linear vs. geometric scales for the dendrogram

F2 - Continue
F1-Help F2-Exit Movement Keys: Home, End, PgUp, PgDn, Up and Down Arrows
  
```

A second help window appears describing the association between ID numbers and the entity names in the data set. Pressing F2 (Continue) removes both help screens and restores the output screen. Press F2 to continue followed by F1 for the help menu, and then press F4 for help on the clustering stages.

```

File: eg.out
  21    3.796E+00    4.160E+00
  22    4.160E+00    4.559E+00
  Help - Clustering stages

This section of the output provides a tabular display of the data
used to develop the dendrogram. At each stage of the clustering, two
clusters are joined (shown in the "Clusters Joined" column) to form
a new cluster. The "Distance" column provide a measure of the
relative similarity of the members of the cluster. The smaller the
distance, the greater the similarity.

F2 - Continue 24 25

F5 - Distances
F6 - Dendrogram
F7 - Linear vs. geometric scales for the dendrogram
F2 - Continue 24 25

F1-Help F2-Exit Movement Keys: Home, End, PgUp, PgDn, Up and Down Arrows

```

Press F2 to continue followed by F1 for the help menu and F5 for help on the distance classes.

```

File: eg.out
  21    3.796E+00    4.160E+00
  22    4.160E+00    4.559E+00
  Help - Distances

The range of relative distance (presented in the output describing
the clustering stages) is divided into 25 discrete classes. This is
necessary to accommodate the techniques used to develop the graphical
depiction of the dendrogram.

F2 - Continue

F4 - Stages of clustering 24 25
F5 - Distances
F6 - Dendrogram
F7 - Linear vs. geometric scales for the dendrogram
F2 - Continue 24 25

F1-Help F2-Exit Movement Keys: Home, End, PgUp, PgDn, Up and Down Arrows

```


Press F2 to continue followed by F1 for the help menu and F6 for the dendrogram help window.

```

File: eg.out
21 3.796E+00 4.160E+00
22 4.160E+00 4.559E+00
Help - Dendrogram

The graphical display from a cluster analysis is referred to as a
dendrogram because of its tree-like appearance. At the "trunk", all
of the entities have been joined into a single cluster (at the far
right of the dendrogram). At the far left, each of the "branches"
represents a single entity and each cluster has only one entity.
Moving from left to right, clusters are joined until all of the
entities have been combined into a single cluster.

The ID values listed along the left margin correspond to those
assigned to the entities in the data set. The values (1-25) along
the top and bottom of the dendrogram correspond to the criterion
values and provide a relative measure of the similarity between
members of a cluster. Clusters at the left are composed of more
similar members than clusters at the right.

F2 - Continue
F1-Help F2-Exit Movement Keys: Home, End, PgUp, PgDn, Up and Down Arrows

```

Press F2 to continue followed by F1 for the help menu and F7 for help on the scales used for depicting the dendrogram.

```

File: eg.out
21 3.796E+00 4.160E+00
22 4.160E+00 4.559E+00
Help - Linear vs. geometric scales

Dendrograms are output using both a linear and geometric scale for
the relative distances between members of a cluster. This is done
because if the range of relative distances is very large, the plot
algorithm gets "confused" when drawing the left side (where the
relative distances are at a minimum) of the dendrogram using a
linear scale. When the range of relative distances is large and a
linear scale is used, there is too much detail on the left side of
the dendrogram for the algorithm to deal with.

When the distance range is large (> than two orders of magnitude)
the dendrogram plotted on a geometric scale will provide a better
representation of the clustering.

F2 - Continue
F2 - Continue
F1-Help F2-Exit Movement Keys: Home, End, PgUp, PgDn, Up and Down Arrows

```

Press F2 to continue and F2 again to exit the program.

Using the dendrogram (the entire output file is presented in the next section) one can see that the two littoral stations (ID numbers 1 and 5) are very similar and are clustered together in the first stage. The inlet station (ID number 4) is very different from all of the other stations and is only grouped with the rest at the last stage. With this information it may be possible to reduce sampling effort at this reservoir by sampling only one of the two littoral stations currently being sampled.

Example Output File

Cluster Analysis

EAU GALLE _____ Title provide in input data set

ID NUMBER	ENTITY	
1	STA10	ID numbers associated with entity names
2	STA20	
3	STA30	
4	STA50	
5	STA60	

Average Linkage Method used for clustering

Method used

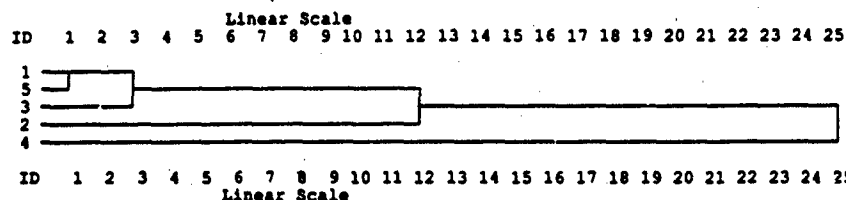
Stage	Clusters Joined		Distance	
1	1	5	6.080E-01	"History" of the clustering
2	1	3	1.219E+00	
3	1	2	3.191E+00	
4	1	4	6.000E+00	

The distances are segmented into the following classes for the Linear dendrogram

CLASS	LOWER BOUND	UPPER BOUND
1	6.080E-01	8.237E-01
2	8.237E-01	1.039E+00
3	1.039E+00	1.255E+00
4	1.255E+00	1.471E+00
5	1.471E+00	1.686E+00
6	1.686E+00	1.902E+00
7	1.902E+00	2.118E+00
8	2.118E+00	2.333E+00
9	2.333E+00	2.549E+00
10	2.549E+00	2.765E+00
11	2.765E+00	2.981E+00
12	2.981E+00	3.196E+00
13	3.196E+00	3.412E+00
14	3.412E+00	3.628E+00
15	3.628E+00	3.843E+00
16	3.843E+00	4.059E+00
17	4.059E+00	4.275E+00
18	4.275E+00	4.490E+00
19	4.490E+00	4.706E+00
20	4.706E+00	4.922E+00
21	4.922E+00	5.137E+00
22	5.137E+00	5.353E+00
23	5.353E+00	5.569E+00
24	5.569E+00	5.784E+00
25	5.784E+00	6.000E+00

The range in distance between the last stage and the first stage of the clustering is divided into 25 equal classes for displaying the dendrogram.

Dendrogram displayed using a linear scale

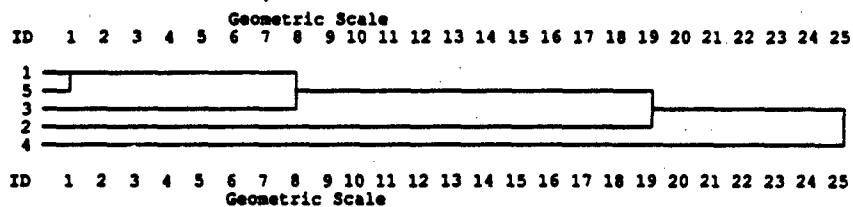


The distances are segmented into the following classes for the Geometric dendrogram

CLASS	LOWER BOUND	UPPER BOUND
1	6.080E-01	6.663E-01
2	6.663E-01	7.302E-01
3	7.302E-01	8.003E-01
4	8.003E-01	8.770E-01
5	8.770E-01	9.611E-01
6	9.611E-01	1.053E+00
7	1.053E+00	1.154E+00
8	1.154E+00	1.265E+00
9	1.265E+00	1.386E+00
10	1.386E+00	1.519E+00
11	1.519E+00	1.665E+00
12	1.665E+00	1.825E+00
13	1.825E+00	2.000E+00
14	2.000E+00	2.191E+00
15	2.191E+00	2.401E+00
16	2.401E+00	2.632E+00
17	2.632E+00	2.884E+00
18	2.884E+00	3.161E+00
19	3.161E+00	3.464E+00
20	3.464E+00	3.796E+00
21	3.796E+00	4.160E+00
22	4.160E+00	4.559E+00
23	4.559E+00	4.996E+00
24	4.996E+00	5.475E+00
25	5.475E+00	6.000E+00

The range in distance between the last stage and the first stage of the clustering is divided into 25 classes using a geometric scale.

Dendrogram displayed using a geometric scale



References

Gaugush, R. F., tech. ed. 1986. Statistical methods for reservoir water quality investigations. Instruction Report E-86-2. Vicksburg, MS: U.S. Army Engineer Waterways Experiment Station.

_____. 1987. Sampling design for reservoir water quality investigations. Instruction Report E-87-1. Vicksburg, MS: U.S. Army Engineer Waterways Experiment Station.

Winer, B. J. 1971. Statistical principles in experimental design. New York: McGraw-Hill.

Bibliography

- Benjamin, J. R., and Cornell, C. A. 1970. *Probability, Statistics, and Decision Making for Civil Engineers*. New York: McGraw-Hill.
- Blum, J. R., and Rosenblatt, J. I. 1972. *Probability and Statistics*. Philadelphia: W. B. Saunders.
- Box, G. E. P., Hunter, W. G., and Hunter, J. S. 1978. *Statistics for Experimenters: An Introduction to Design, Data Analysis, and Model Building*. New York: John Wiley and Sons.
- Cochran, W. G. 1977. *Sampling Techniques*. New York: John Wiley and Sons.
- Cochran, W. G., and Cox, G. M. 1957. *Experimental Designs*. New York: John Wiley and Sons.
- Green, R. H. 1979. *Sampling Design and Statistical Methods for Environmental Biologists*. New York: John Wiley and Sons.
- Snedecor, G. W., and Cochran, W. G. 1967. *Statistical Methods*. Ames, IA: Iowa State University Press.
- Sokal, R. R., and Rohlf, F. J. 1969. *Biometry*. San Francisco: W. H. Freeman.
- Steel, R. G. D., and Torrie, J. H. 1980. *Principles and Procedures in Statistics*. 2nd ed. New York: McGraw-Hill.
- Stuart, A. 1962. *Basic Ideas of Scientific Sampling*. New York: Hafner Publishing.
- Tukey, J. W. 1977. *Exploratory Data Analysis*. Reading, MA: Addison-Wesley.
- Wine, R. L. 1964. *Statistics for Scientists and Engineers*. Englewood Cliffs, NJ: Prentice-Hall.

Wonnacott, T. H., and Wonnacott, R. J. 1972. *Introductory Statistics*.
New York: John Wiley and Sons.

Zar, J. H. 1974. *Biostatistical Analysis*. Englewood Cliffs, NJ:
Prentice-Hall.

REPORT DOCUMENTATION PAGEForm Approved
OMB No. 0704-0188

Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Washington Headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302, and to the Office of Management and Budget, Paperwork Reduction Project (0704-0188), Washington, DC 20503.

1. AGENCY USE ONLY (Leave blank)

2. REPORT DATE

February 1993

3. REPORT TYPE AND DATES COVERED

Final report

4. TITLE AND SUBTITLE

Sampling Design Software User's Manual

5. FUNDING NUMBERS

6. AUTHOR(S)

Robert F. Gaugush

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

U.S. Army Engineer Waterways Experiment Station
Environmental Laboratory
3909 Halls Ferry Road, Vicksburg, MS 39180-61998. PERFORMING ORGANIZATION
REPORT NUMBER

Instruction Report W-93-1

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)

U.S. Army Corps of Engineers, Washington, DC 20314-1000

10. SPONSORING / MONITORING
AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES

Available from National Technical Information Service, 5285 Port Royal Road, Springfield, VA 22161.
Sampling Design Software (SDS, Version 2.0) is provided with manual.

12a. DISTRIBUTION / AVAILABILITY STATEMENT

Approved for public release; distribution is unlimited.

12b. DISTRIBUTION CODE

13. ABSTRACT (Maximum 200 words)

Sampling Design Software consists of four interactive programs (DECMATRX, VARCOM, ERROR, and CLUSTER) that assist users in the development of water quality sampling programs. The user's manual describes data entry and program use and provides examples. Example screen displays are also provided.

14. SUBJECT TERMS

Cluster analysis
Decision matrices
Error analysisSample design
Variance component analysis
Water quality

15. NUMBER OF PAGES

70

16. PRICE CODE

17. SECURITY CLASSIFICATION
OF REPORT

UNCLASSIFIED

18. SECURITY CLASSIFICATION
OF THIS PAGE

UNCLASSIFIED

19. SECURITY CLASSIFICATION
OF ABSTRACT

20. LIMITATION OF ABSTRACT